Segmentez les clients d'un site e-commerce

Synthèse

Contexte & Objectifs

PROJET

Sur la base des transactions effectuées sur 2 années pleines (septembre 2016 à août 2018), établir une segmentation des clients permettant de décrire leur comportement d'achat à des fins de marketing. Le nombre de classes devra être suffisamment élevé pour permettre une stratégie diversifiée en fonction des segments, tout en n'étant pas excessif pour rester interprétable dans le contexte métier.

DÉMARCHE

Après une analyse exploratoire des données, plusieurs modèles de classification non supervisée (clustering) seront testés et évalués en fonction de leur capacité à répondre à la problématique métier.

olist

Sommaire

olist

1 – Analyse Exploratoire des Données

1.1 – Exploration & Nettoyage

JEU DE DONNEES INITIAL* :

5 tables de dimensions

Données géographiques – table 'geolocation'
1,000,163 lignes x 6 colonnes – 0% NaN

Clients – table 'customers'
99,441 lignes x 2 colonnes – 0% NaN

Vendeurs – table 'sellers'
3,95 lignes x 2 colonnes – 0% NaN

Description produits – table 'products'
32,951 lignes x 9 colonnes – 0.8% NaN

Catégories produits – table 'categories'
71 lignes x 2 colonnes – 0% NaN

4 tables de faits

Ventes – table 'baskets'
112,650 lignes x 7 colonnes – 0% NaN

Paiements – table 'payments'
103,886 lignes x 5 colonnes – 0% NaN

Revues de produit – table 'reviews'
99,224 lignes x 7 colonnes – 21% NaN

Livraisons – table 'shipping'
99,441 lignes x 8 colonnes – 0.6% NaN

olist

* Source : https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Voir Annexe 2 pour le schéma de la base de données SQL.

1 – Analyse Exploratoire des Données

1.2 – Feature Engineering & Sélection

Feature engineering

- Longueur des titres et reviews
- Distance entre clients et vendeurs
- Groupement des produits en 13 puis 3 catégories
- Labels pour les transactions faites le weekend ou le soir après 18h
- Nombre de jours depuis la dernière transaction du client
- Ajout de données externes etc…*

Retraitements

- Suppressions des valeurs manquantes & doublons
- Suppression des variables trop corrélées entre elles
- Jointures entre les tables
- Agrégation des données par client
- Création itérative de plusieurs jeux de données avec différentes features numériques & catégorielles

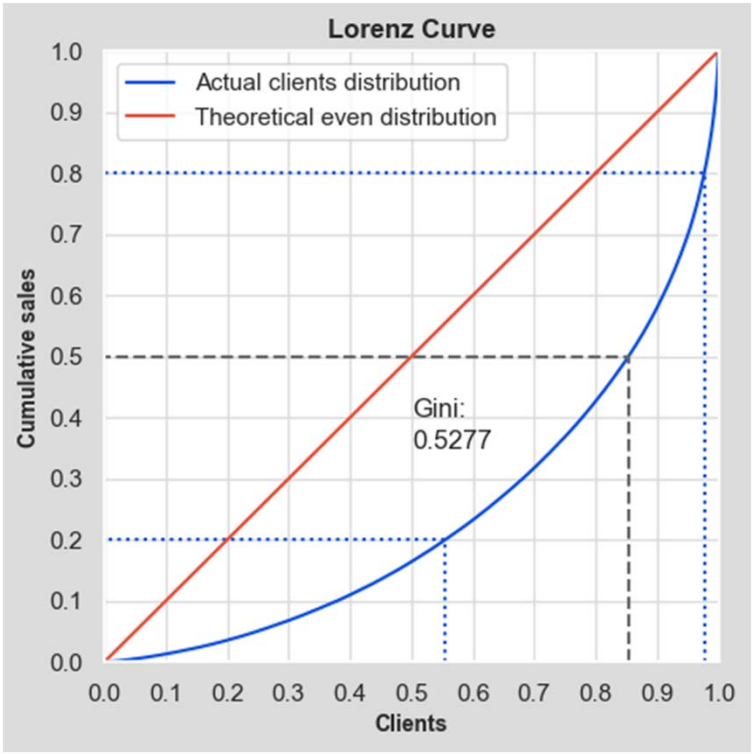
olist

*Voir section 2.2 du notebook 1 (livrable 2) pour la liste complète des features ajoutées / modifiées

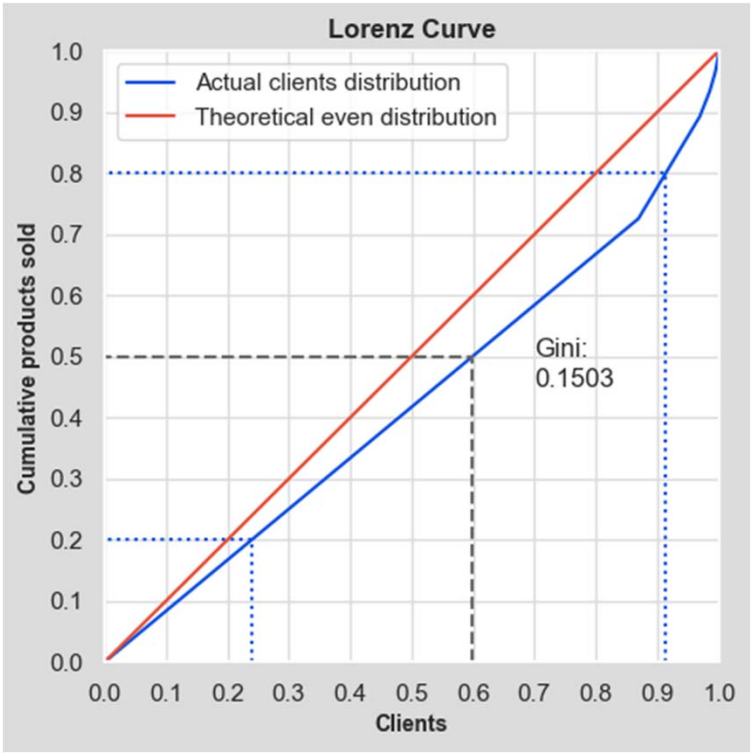
1 – Analyse Exploratoire des Données

1.3 – Statistiques descriptives – Chiffre d'affaires en Valeur & Volume

Valeur - Chiffre d'affaires total :

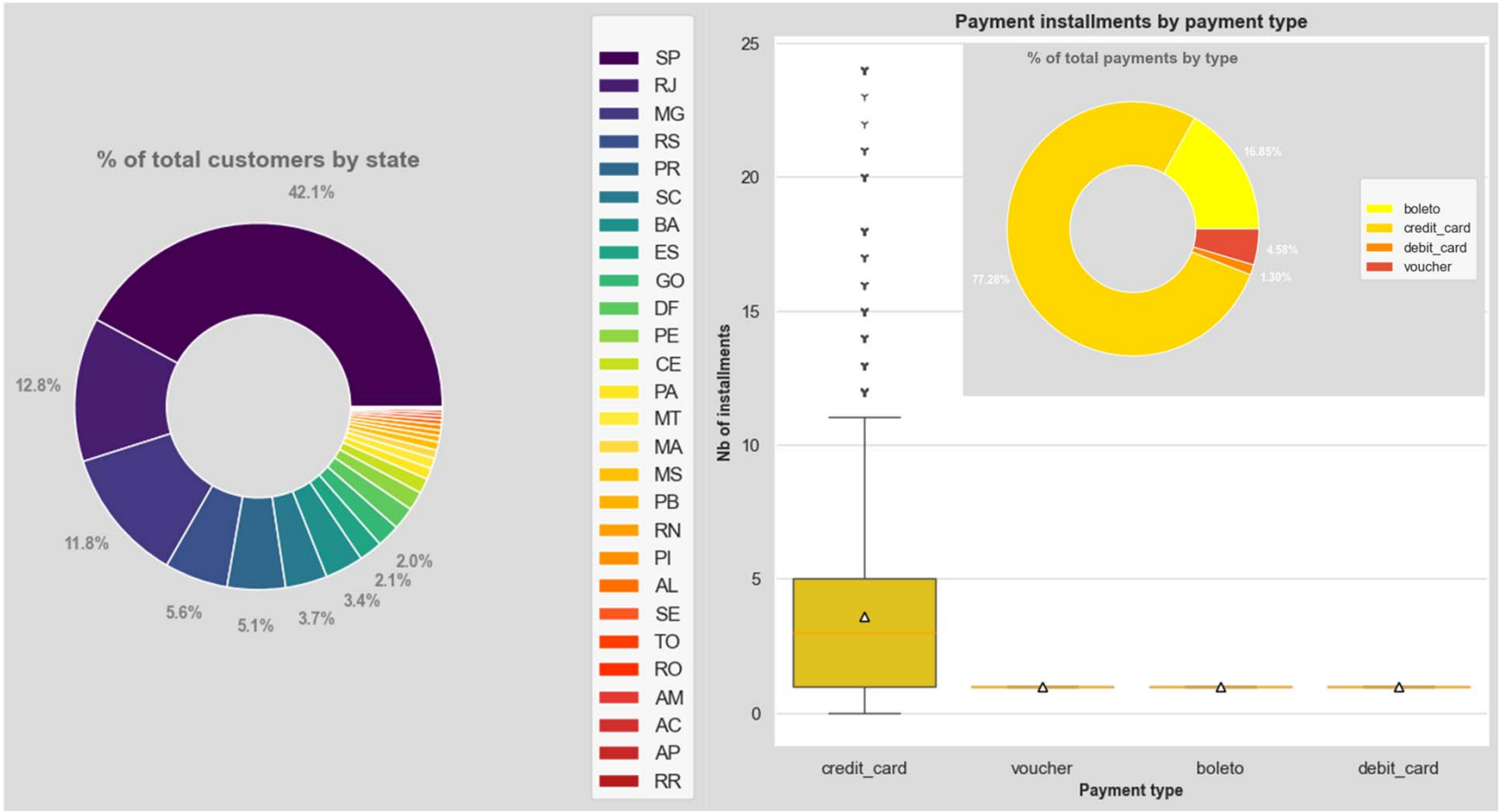


Volume - Nombre de produits vendus :



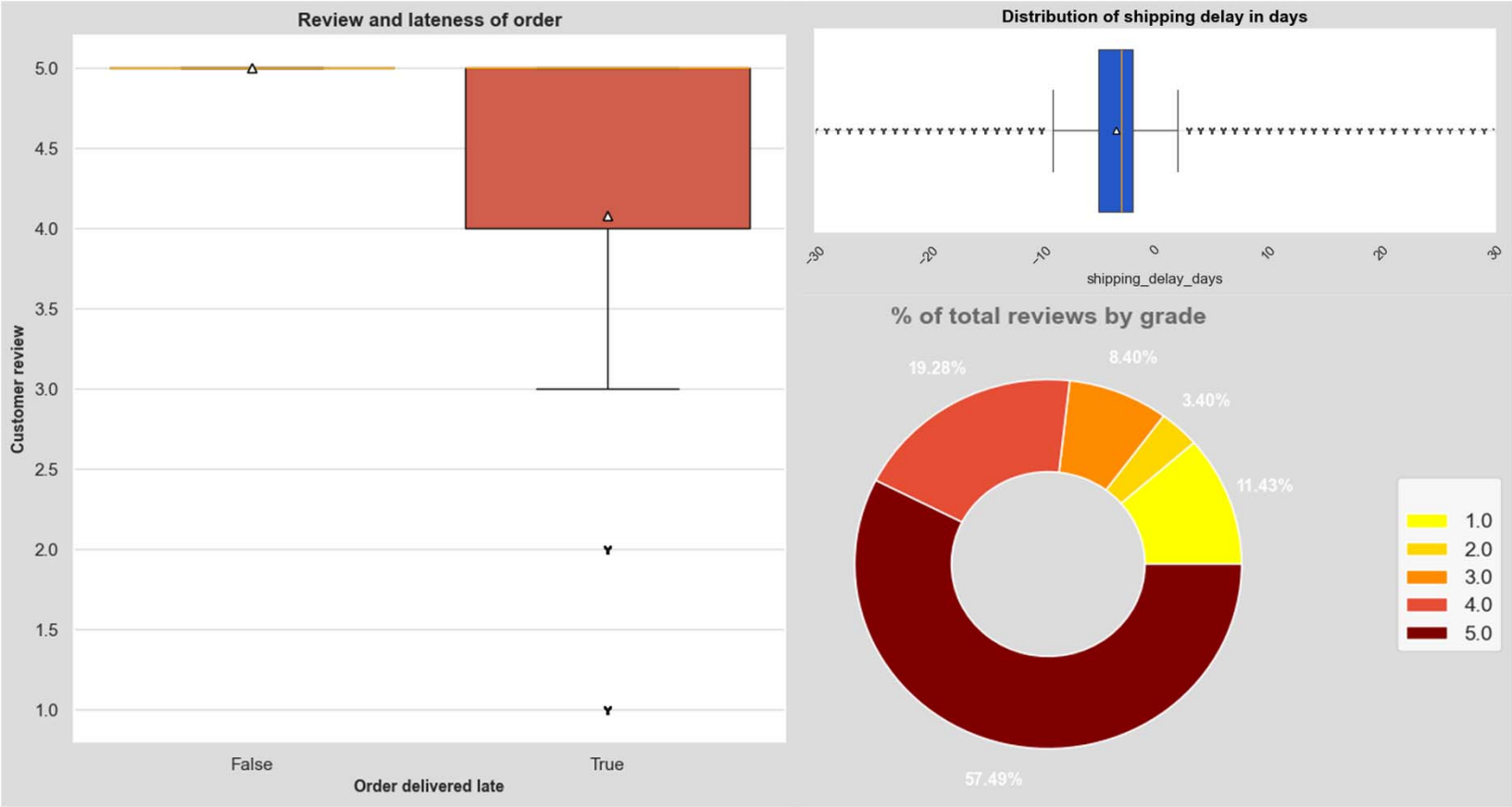
1 – Analyse Exploratoire des Données

1.3 – Statistiques descriptives – Localisation & Types de Paiements



1 – Analyse Exploratoire des Données

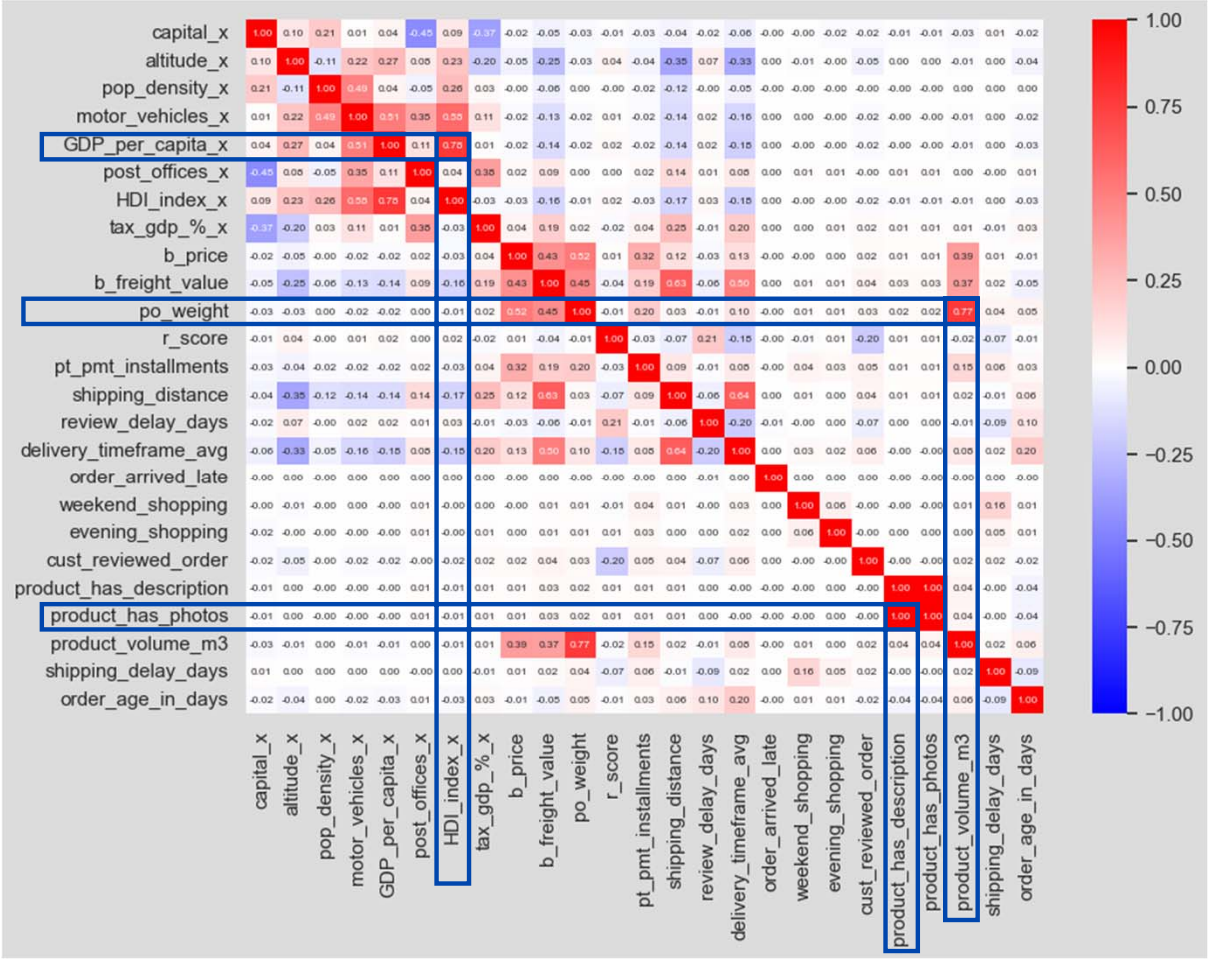
1.3 – Statistiques descriptives – Reviews & Retards d'Acheminement



1 – Analyse Exploratoire des Données

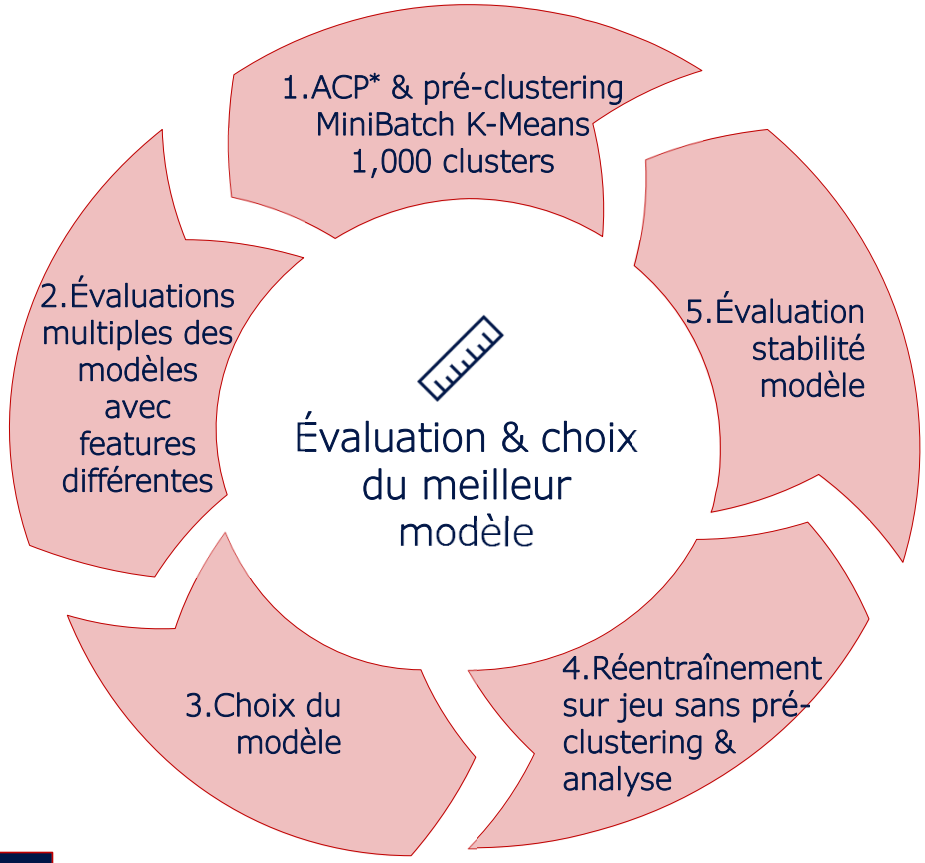
1.3 – Statistiques descriptives - Corrélations

- Suppression des features fortement corrélées
- Skew
- Variables non-gaussiennes
  - Coefficient de Spearman



2 – Modélisation

2.1 – Problématique & Méthodologie



*Voir Annexe 6 pour les résultats de l'ACP

2. Métriques d'Évaluation du Nombre de Clusters Optimal:

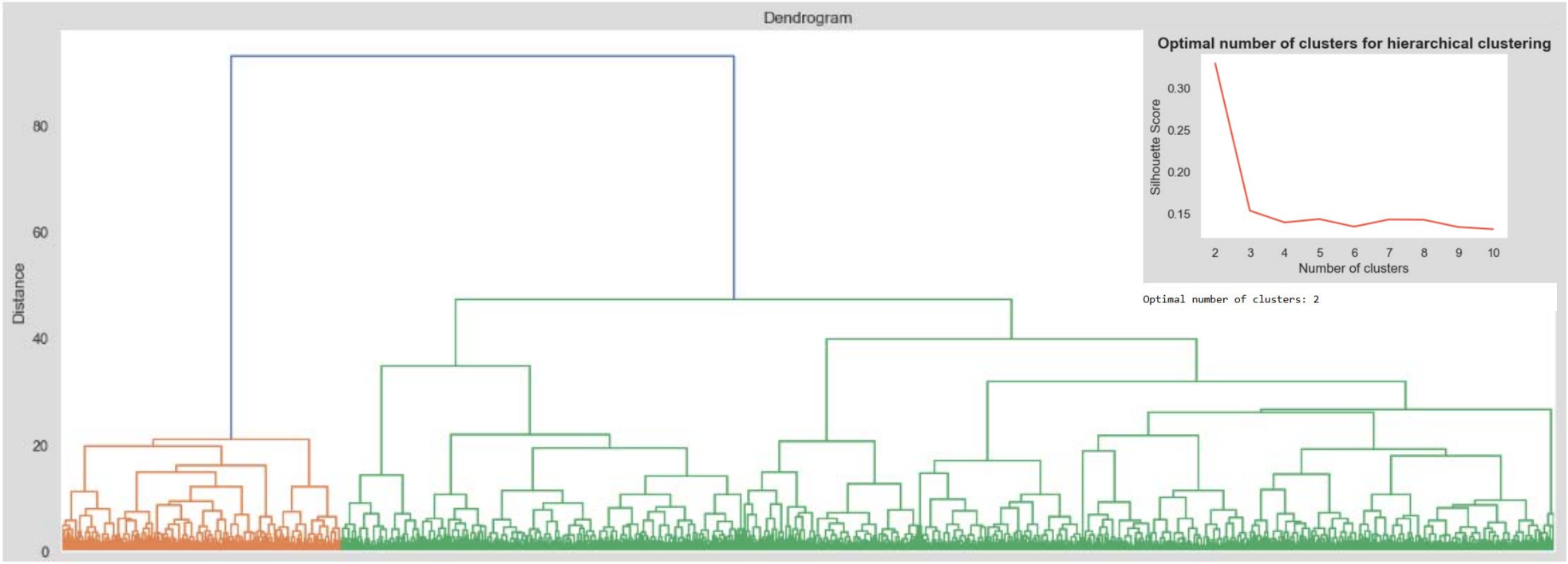
- Méthode du 'coude'
- Coefficient Silhouette
- Indice de Davies-Bouldin
- Indice de Kalinski-Harabasz

3. Modèles Testés:

- Classification Ascendante Hiérarchique
- DBSCAN
- HDBSCAN
- Affinity Propagation
- K-Prototypes
- K-means

2 – Modélisation

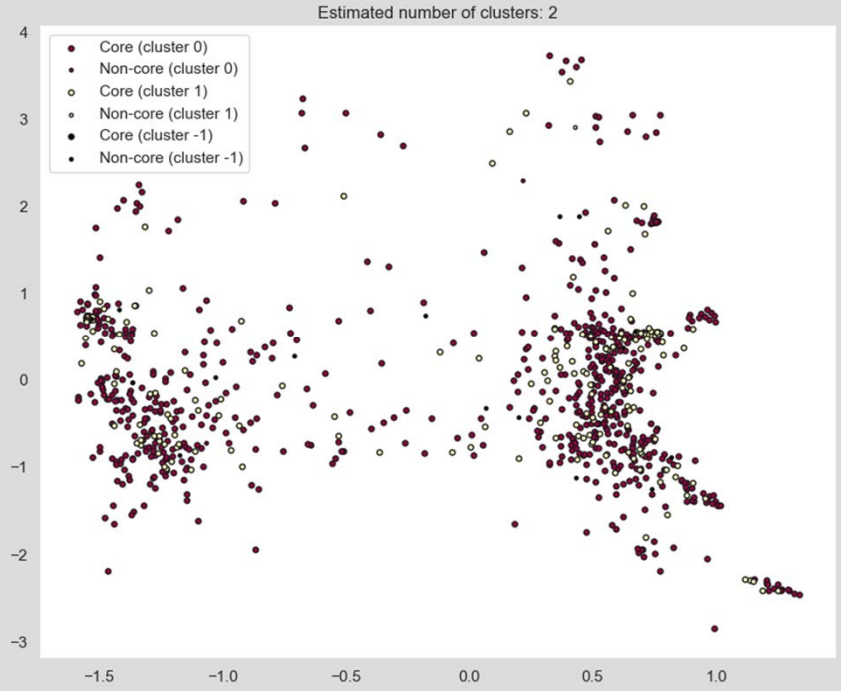
2.2 – Critères Métier & Choix du Modèle Final – Classification Ascendante Hiérarchique

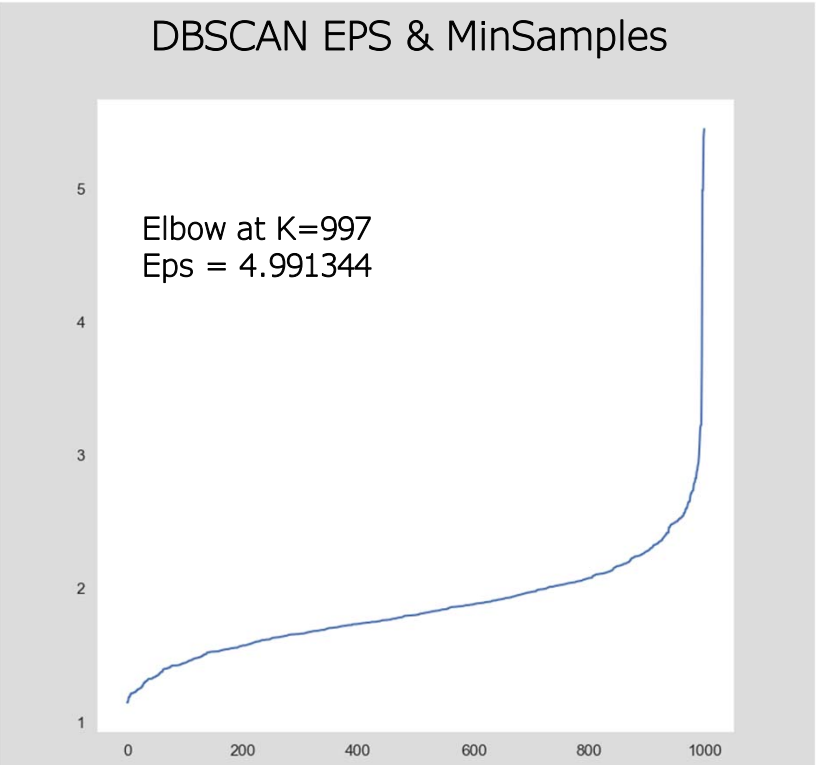



```
Nb of customer pre-clusters in 1st cluster : 187 . Population :  12437
Nb of customer pre-clusters in 2nd cluster : 813 . Population :  82903
Total customers : 95340 . Total pre-clusters : 1000
```

2 – Modélisation

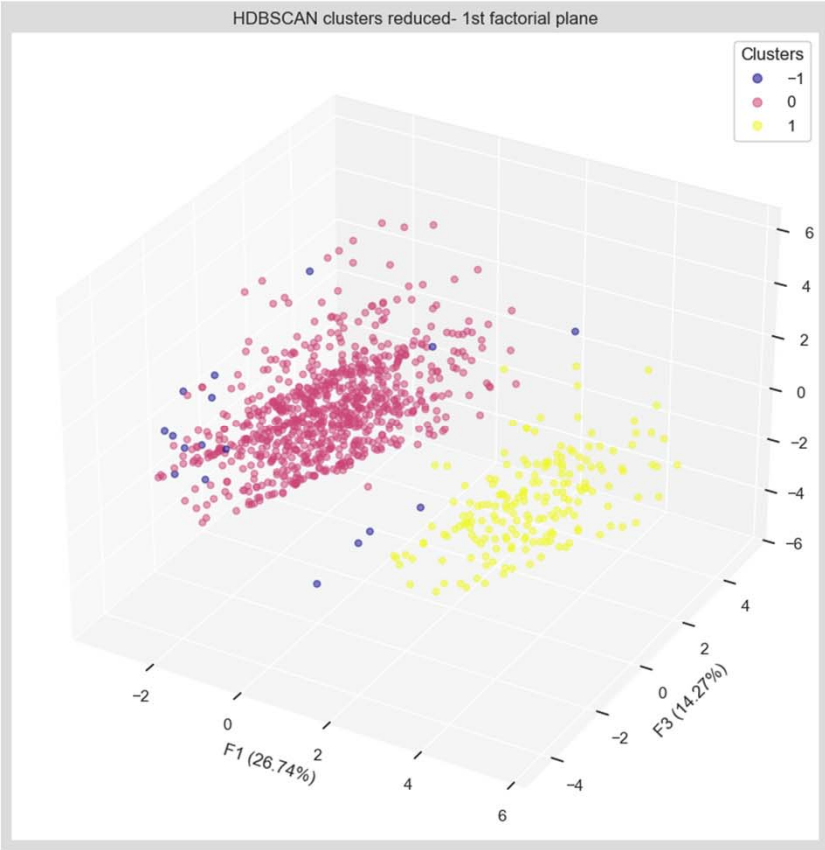
2.2 – Critères Métier & Choix du Modèle Final - DBSCAN

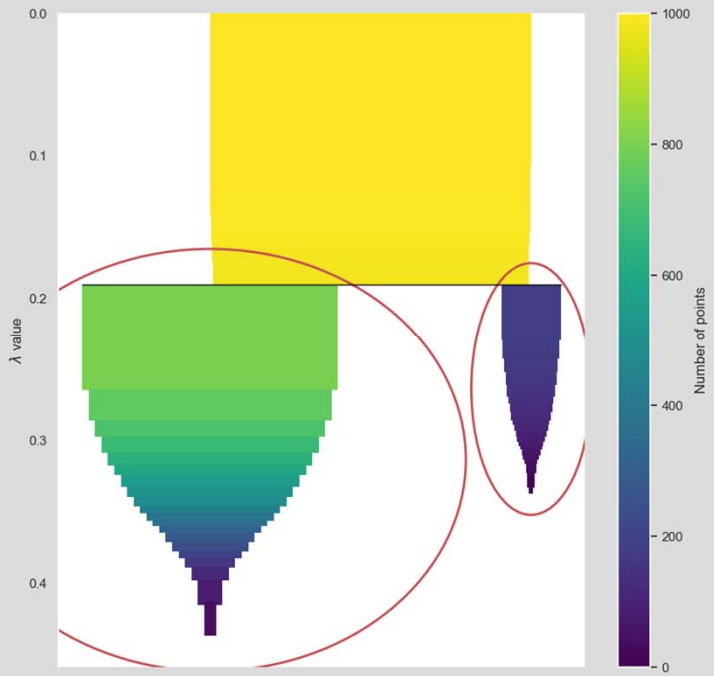




2 – Modélisation

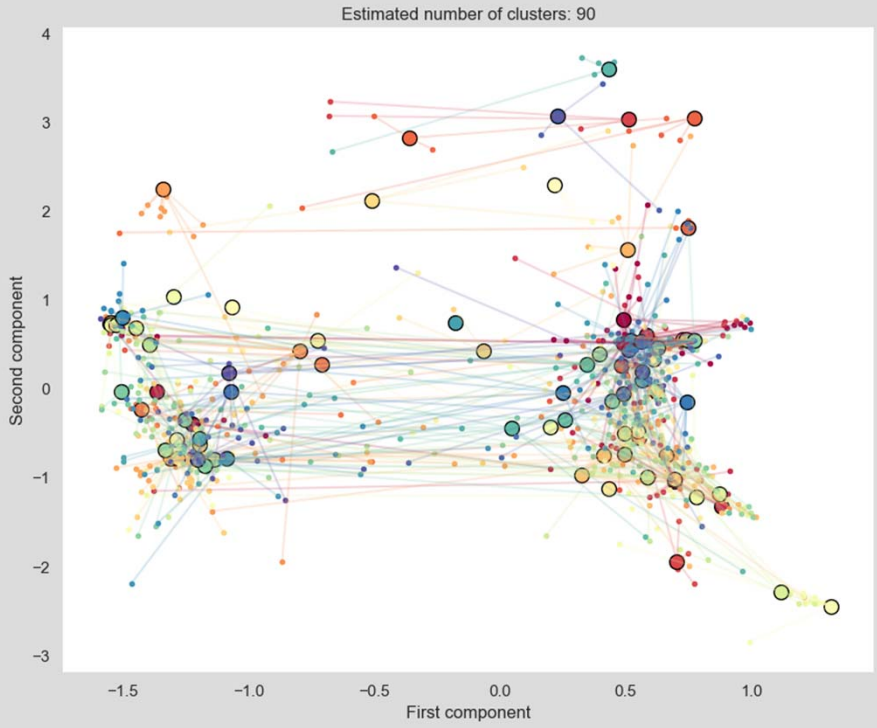
2.2 – Critères Métier & Choix du Modèle Final - HDBSCAN

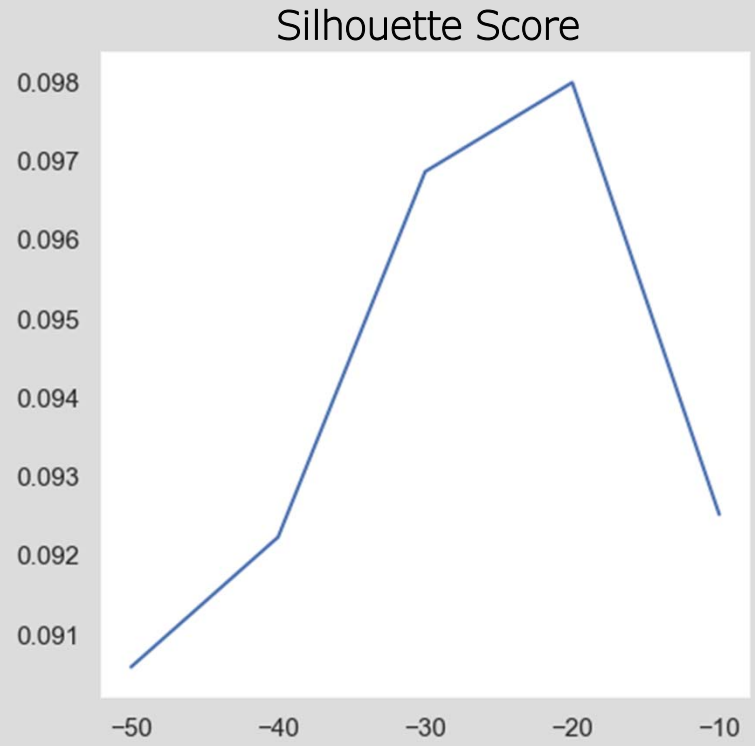




2 – Modélisation

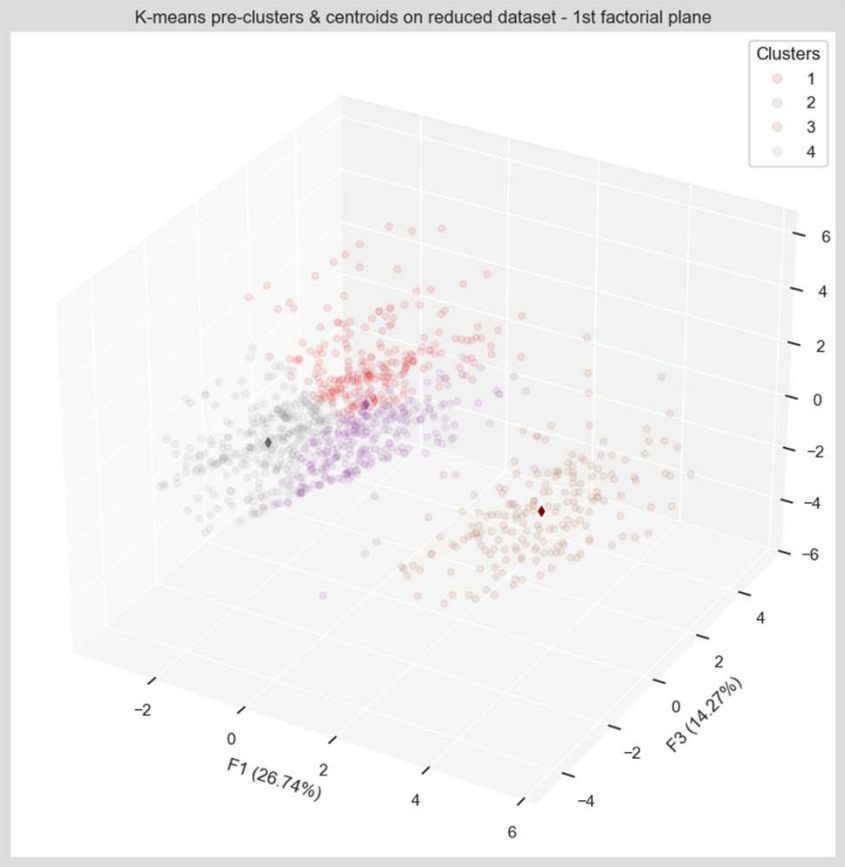
2.2 – Critères Métier & Choix du Modèle Final – Affinity Propagation

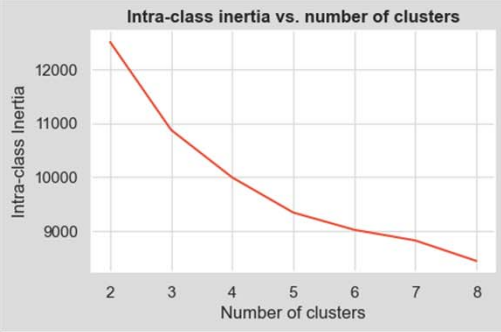


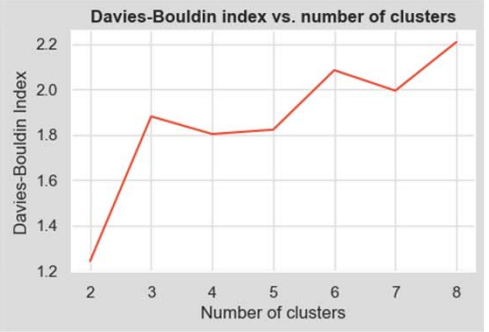


2 – Modélisation

2.2 – Critères Métier & Choix du Modèle Final – K-Means







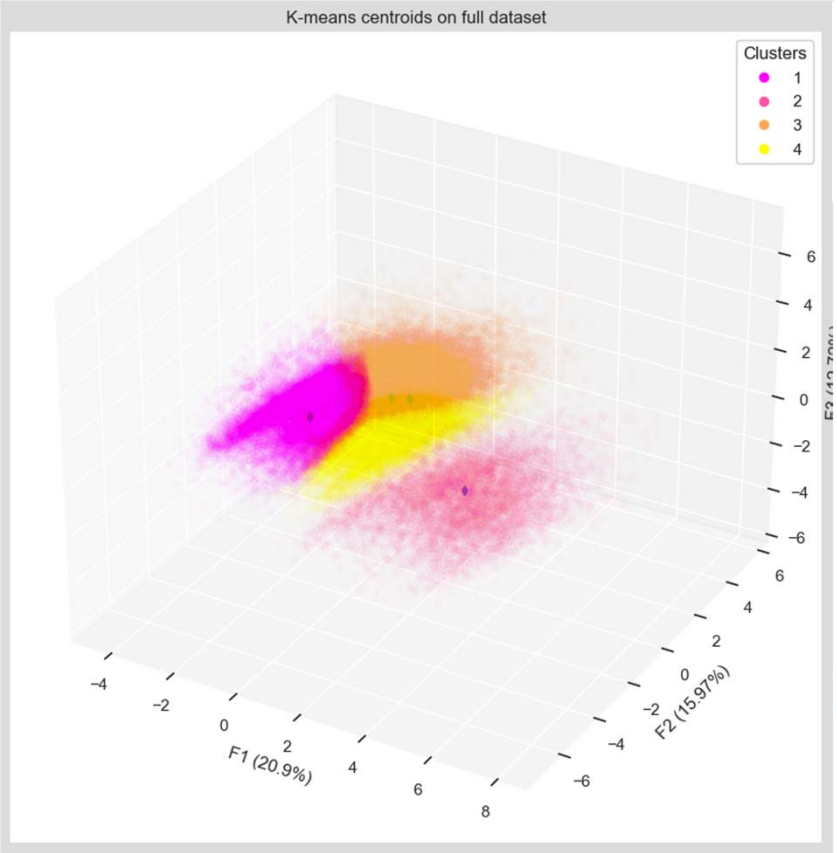

```
Nb of customer pre-clusters in 1st cluster : 226 . Population :  19771
Nb of customer pre-clusters in 2nd cluster : 279 . Population :  27386
Nb of customer pre-clusters in 3rd cluster : 189 . Population :  12481
Nb of customer pre-clusters in 4th cluster : 306 . Population :  35702
Total customers : 95340 . Total pre-clusters : 1000
```

<u>Meilleur modèle - compromis :</u>

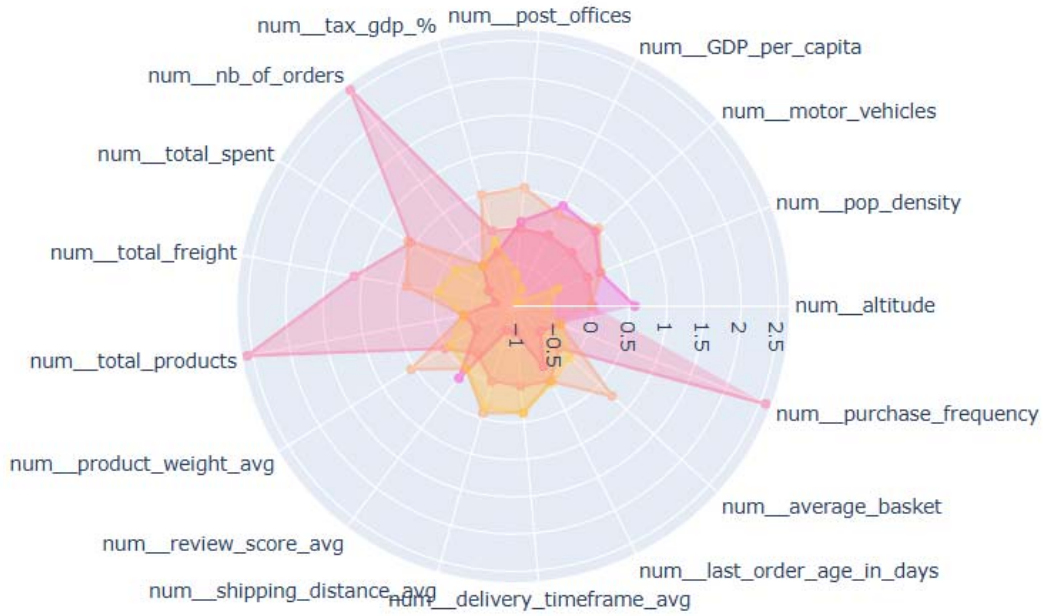
- Nombre adéquat de clusters pour une stratégie marketing différenciée
- Facilement lisible & interprétable par le métier

2 – Modélisation

2.2 – Critères Métier & Choix du Modèle Final – K-Means (sans Pré-Clustering)

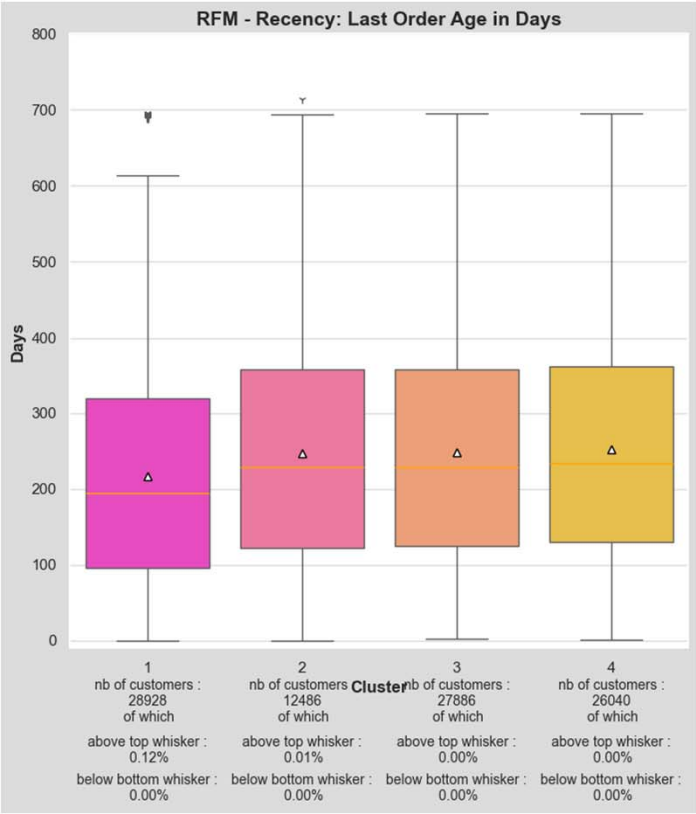


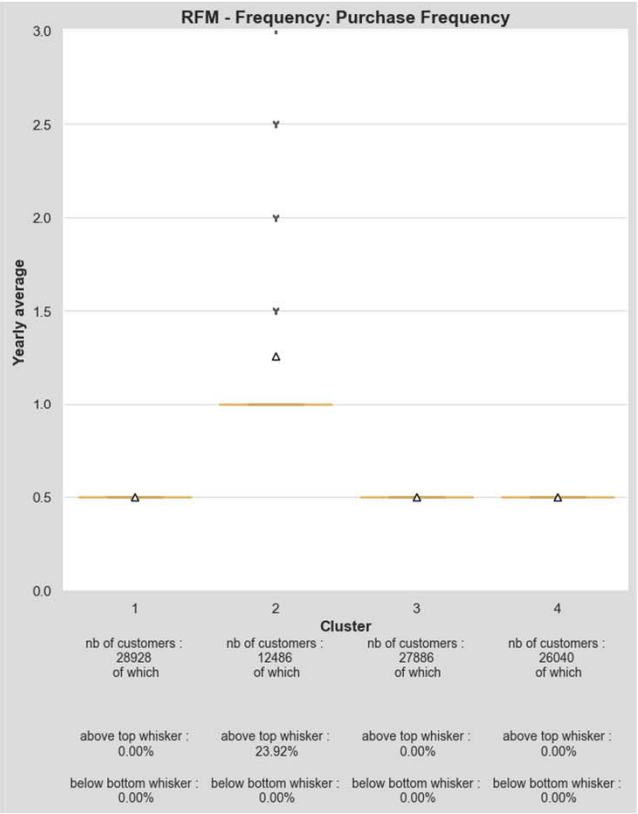

```
Nb of customers in 1st cluster : 28928
Nb of customers in 2nd cluster : 12486
Nb of customers in 3rd cluster : 27886
Nb of customers in 4th cluster : 26040
Total customers : 95340
```



2 – Modélisation

2.3 – Caractéristiques RFM* des Clusters & Stratégie Marketing – Récence & Fréquence





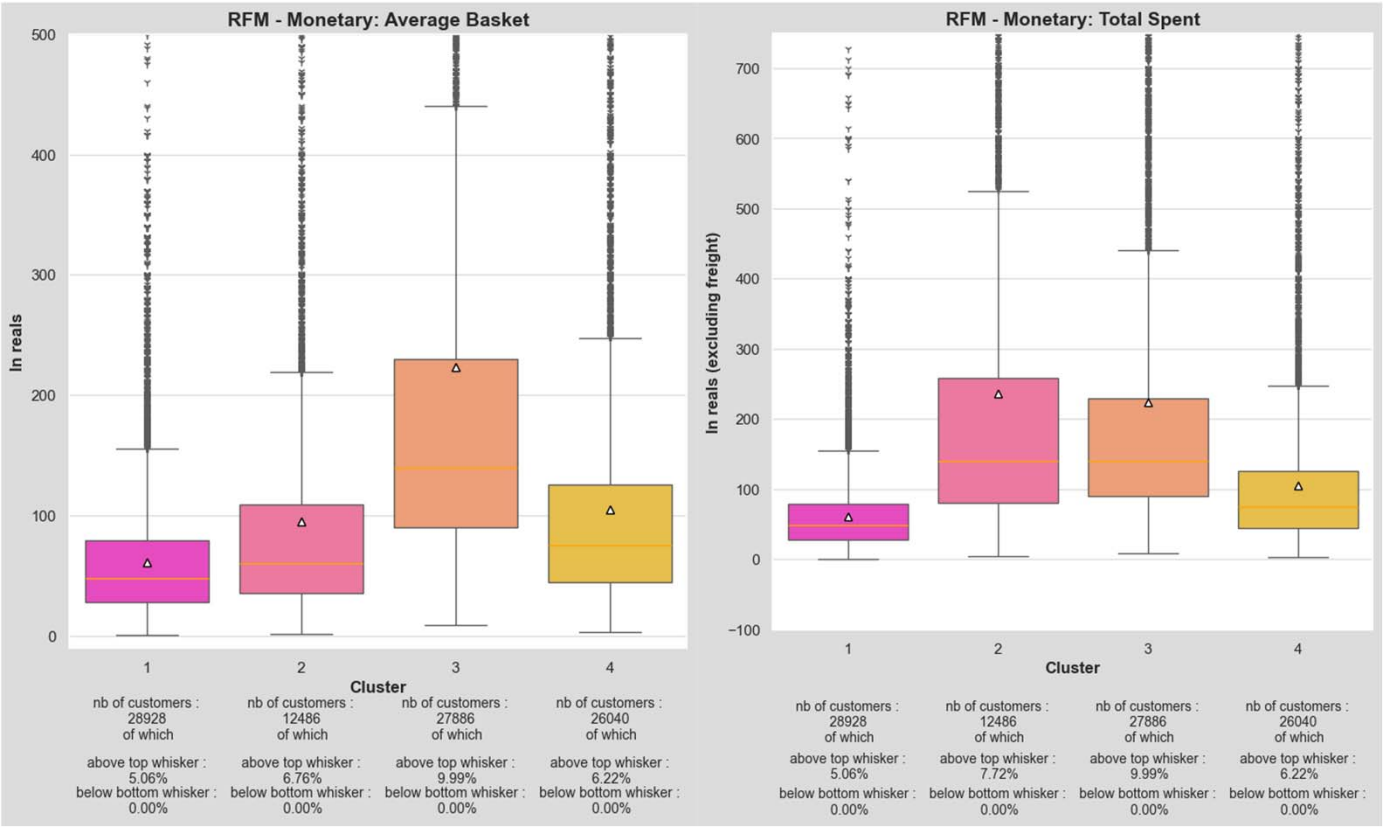
Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives sauf:

- <u>Récence :</u> clusters 2 et 3
- <u>Fréquence :</u> seul le cluster 2 diffère significativement

*Voir Annexes 4A et 4B pour les autres caractéristiques des clusters

2 – Modélisation

2.3 – Caractéristiques RFM* des Clusters & Stratégie Marketing – Montant 1/2



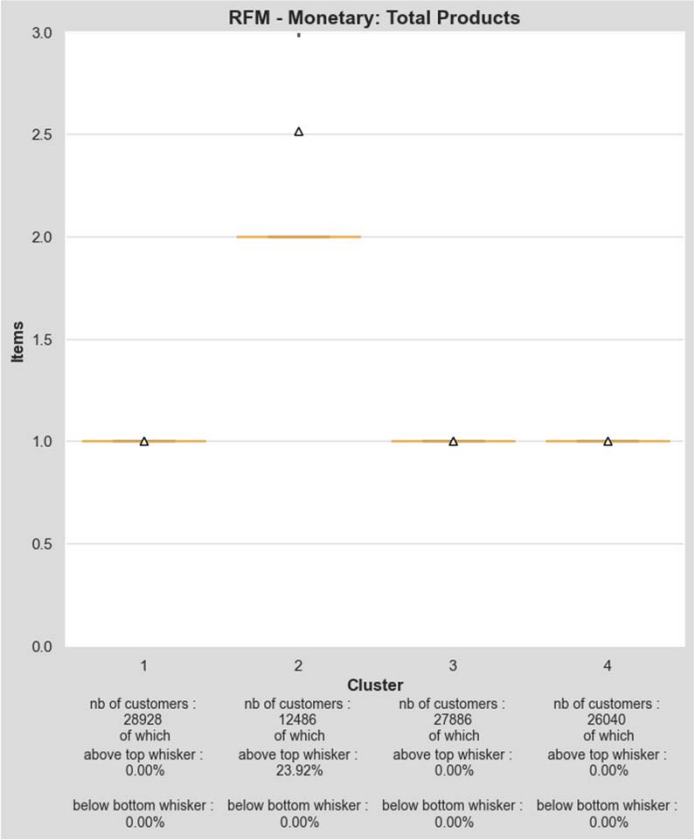
Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives entre tous les clusters sur:

- Panier moyen
- Dépense totale

*Voir Annexes 4A et 4B pour les autres caractéristiques des clusters

2 – Modélisation

2.3 – Caractéristiques RFM* des Clusters & Stratégie Marketing – Montant 2/2



Tests de Kruskal-Wallis & Dunn => différences non statistiquement significatives entre tous les clusters sauf Cluster 2 pour le nombre total de produits achetés

*Voir Annexes 4A et 4B pour les autres caractéristiques des clusters

2 – Modélisation

2.3 – Caractéristiques RFM* des Clusters & Stratégie Marketing – Synthèse*

Cluster 1 - chenilles

- Clients peu engagés : transactions plus récentes mais peu fréquentes et de faible valeur monétaire
- Stratégie : offres d'entrée de gamme
- Priorité : basse

Cluster 2 - abeilles

- Clients très engagés : transactions moins récentes mais plus nombreuses et plus élevées en montant
- Stratégie : offres personnalisées
- Priorité : haute

Cluster 3 - coccinelles

- Clients sélectifs : transactions moins récentes mais panier moyen et dépense totale les plus élevés
- Stratégie : réduction sur achats en volume ou répétés
- Priorité : moyenne

Cluster 4 – papillons

- Clients testeurs : transactions plus anciennes et de valeur moyenne
- Stratégie : programme de fidélisation
- Priorité : moyenne

*Voir Annexes 4A et 4B pour les autres caractéristiques des clusters

3 – Evaluation du modèle Final

3.1 – Stabilité des Clusters à l'Initialisation

Paramètres:

- 20 itérations / modèle
- random_state = indice itération
- Calcul ARI / itération => moyenne

Résultats & Interpretation:

- MiniBatch K-Means
  - ARI_{MOY} = 0.4788
  - Variabilité accrue – échantillons aléatoires & compromis vitesse/précision (taille des échantillons)

- K-Means
  - ARI_{MOY} = 0.8360
  - Bonne stabilité à l'initialisation & concordance des clusters entre 2 itérations

olist

3 – Évaluation du modèle Final

3.2 – Contrat de Maintenance – Méthodologie

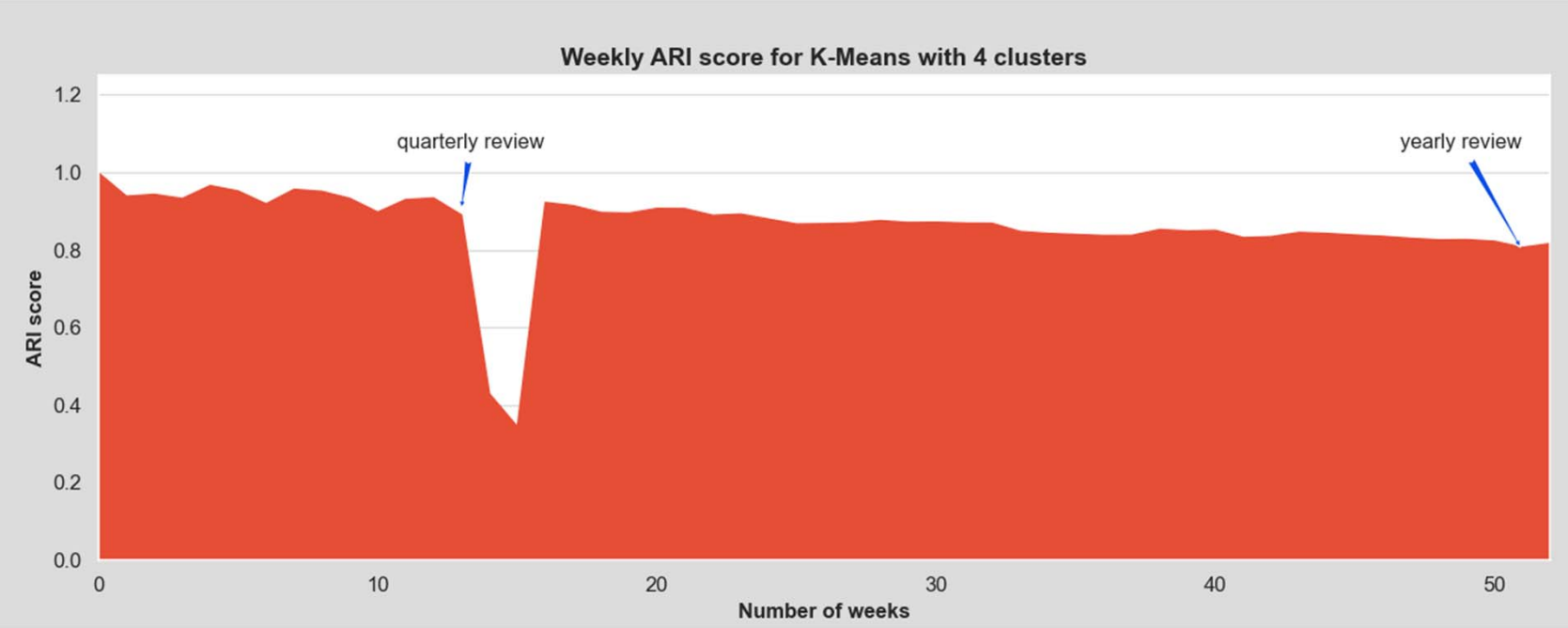
Modèle =>K-Means (K=4)

- T_0 = date dernière transaction client placée (29/08/2018)
  - Entraînement du modèle M_0 à T_0 sur les données D_0

- T_1,…, T_{52} = dates 1, … , 52 semaines avant T_0
  - Génération d'un fichier de données D_i contenant toutes les transactions depuis le début de l'historique disponible (15/09/2016) jusqu'à D_i
  - Prédiction des clusters C_{0i} du modèle M_0 à T_i sur les données D_i
  - Réentraînement du modèle M_i à T_i sur les données D_i et prédiction des clusters C_{ii}
  - Calcul de l'ARI entre C_{0i} & C_{ii}

3 – Évaluation du modèle Final

3.2 – Contrat de Maintenance – Recommandations



Conclusion & perspectives

- Limites de la méthodologie RFM
  - Récence & fréquence inadaptées à l'analyse de transactions uniques
  - Ne mesure pas le « churn risk » ni la loyauté
    - à la "Marketplace" ou à la marque/produit acheté?
  - Image incomplète des comportements d'achat qui ignore:
    - la saisonnalité
    - les facteurs psychologiques
    - les influences socio-économico-démographiques

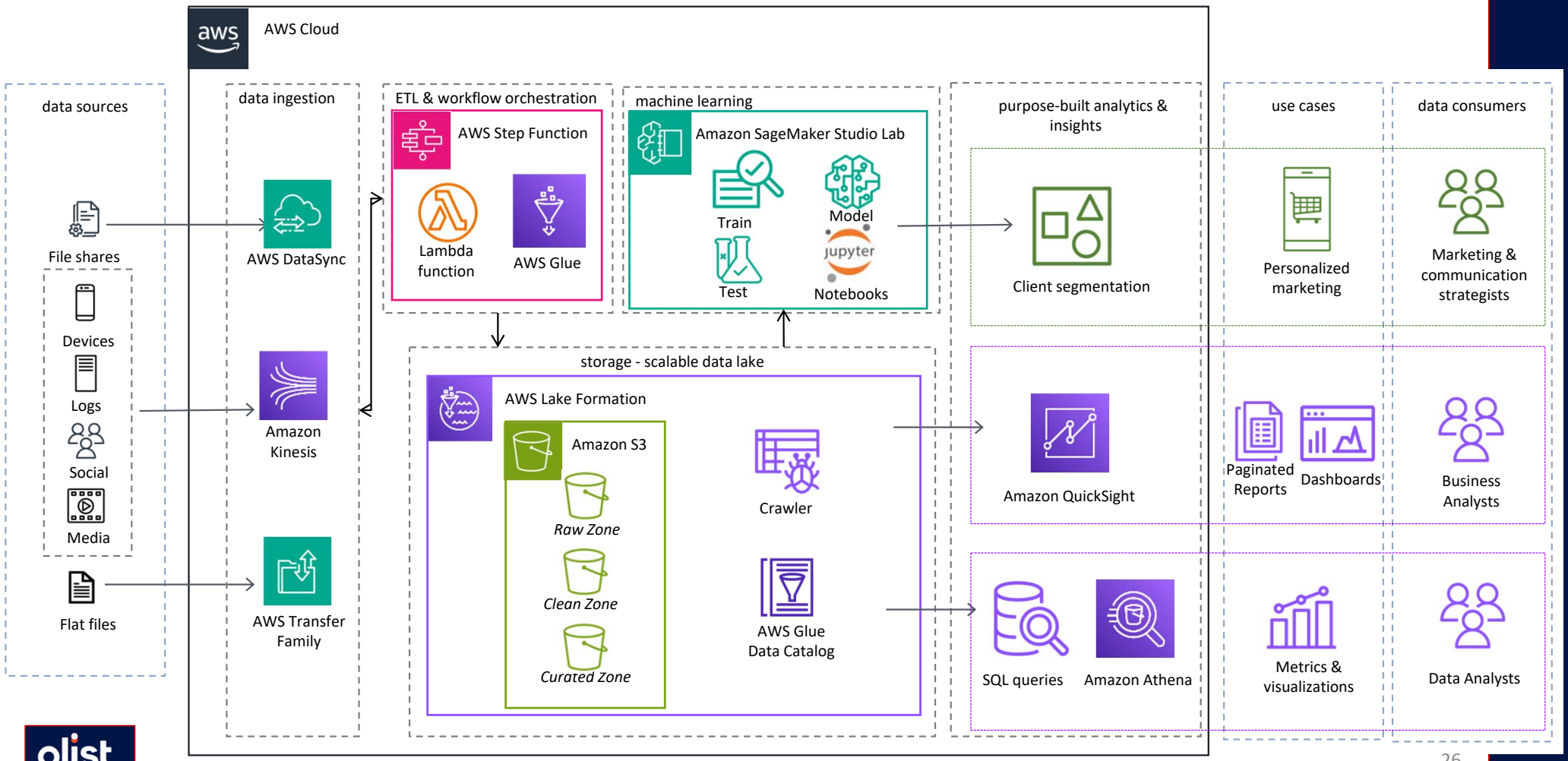
- Affiner la stratégie
  - Identifier les clients potentiels & intégrer les coûts d'acquisition
  - Enrichissement des données (démographiques, économiques etc…)*
  - Attribution par canal & « marketing-mix »

olist

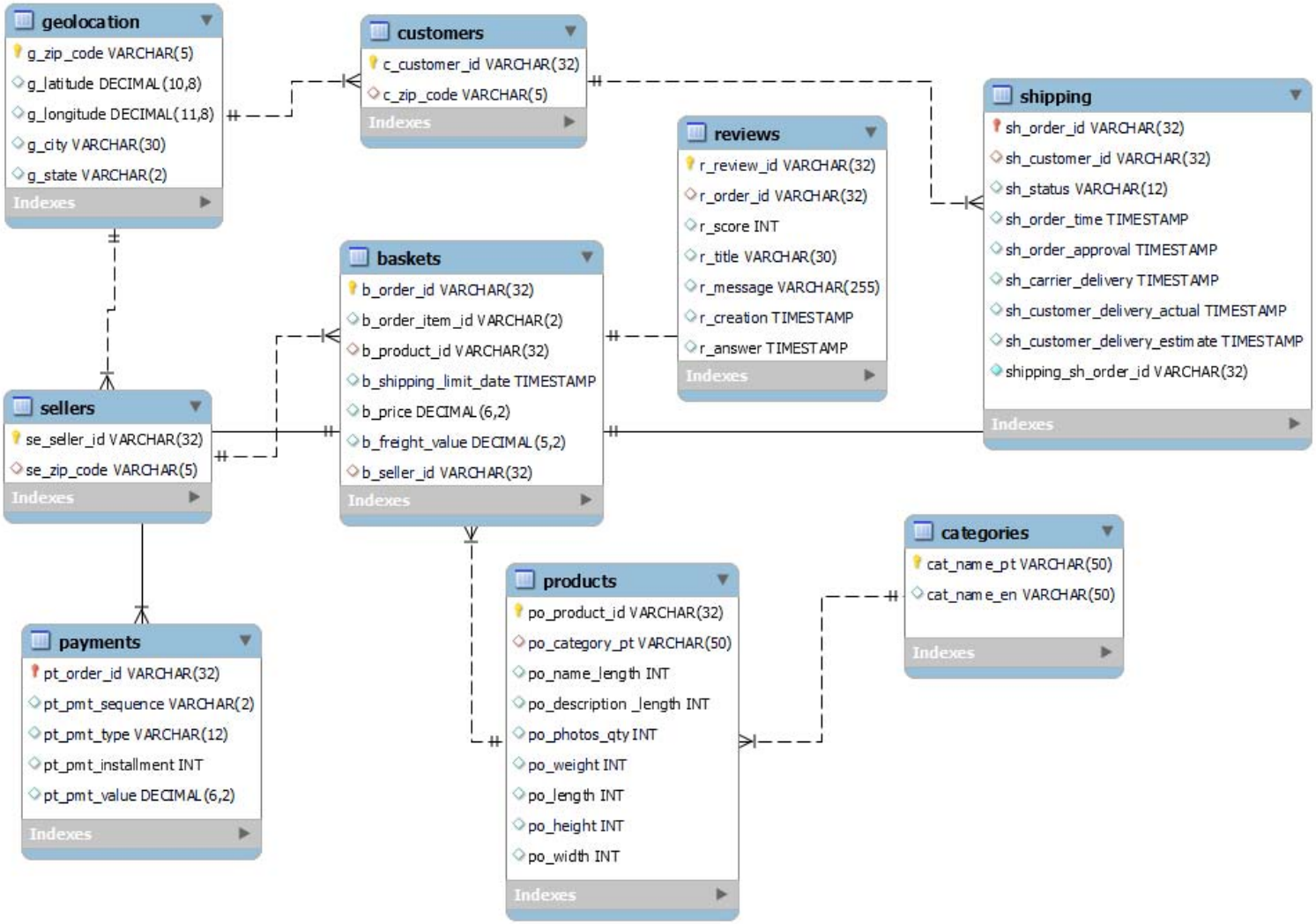
*Voir exemples en Annexes 4A et 4B

ANNEXES

ANNEXE 1 – ARCHITECTURE AWS DONNÉES, MODÉLISATION & REPORTING



ANNEXE 2 – ERD


geolocation
g_zip_code VARCHAR(5)
g_latitude DECIMAL(10,8)
g_longitude DECIMAL(11,8)
g_city VARCHAR(30)
g_state VARCHAR(2)
Indexes
customers
c_customer_id VARCHAR(32)
c_zip_code VARCHAR(5)
Indexes
reviews
r_review_id VARCHAR(32)
r_order_id VARCHAR(32)
r_score INT
r_title VARCHAR(30)
r_message VARCHAR(255)
r_creation TIMESTAMP
r_answer TIMESTAMP
Indexes
shipping
sh_order_id VARCHAR(32)
sh_customer_id VARCHAR(32)
sh_status VARCHAR(12)
sh_order_time TIMESTAMP
sh_order_approval TIMESTAMP
sh_carrier_delivery TIMESTAMP
sh_customer_delivery_actual TIMESTAMP
sh_customer_delivery_estimate TIMESTAMP
shipping_sh_order_id VARCHAR(32)
Indexes
baskets
b_order_id VARCHAR(32)
b_order_item_id VARCHAR(2)
b_product_id VARCHAR(32)
b_shipping_limit_date TIMESTAMP
b_price DECIMAL(6,2)
b_freight_value DECIMAL(5,2)
b_seller_id VARCHAR(32)
Indexes
sellers
se_seller_id VARCHAR(32)
se_zip_code VARCHAR(5)
Indexes
payments
pt_order_id VARCHAR(32)
pt_pmt_sequence VARCHAR(2)
pt_pmt_type VARCHAR(12)
pt_pmt_installment INT
pt_pmt_value DECIMAL(6,2)
Indexes
products
po_product_id VARCHAR(32)
po_category_pt VARCHAR(50)
po_name_length INT
po_description _length INT
po_photos_qty INT
po_weight INT
po_length INT
po_height INT
po_width INT
Indexes
categories
cat_name_pt VARCHAR(50)
cat_name_en VARCHAR(50)
Indexes

ANNEXE 3 – REQUÊTES SQL / KPI DASHBOARD

3.2%

des commandes sur les 3 derniers mois ont été livrées au client avec au moins

3 jours de retard

0.5%

des vendeurs ont généré un CA

> 100,000 réals

sur les commandes livrées

Sur les 1,495 nouveaux vendeurs dans les 3 derniers mois, seuls

2

ont vendu

plus de 30 produits

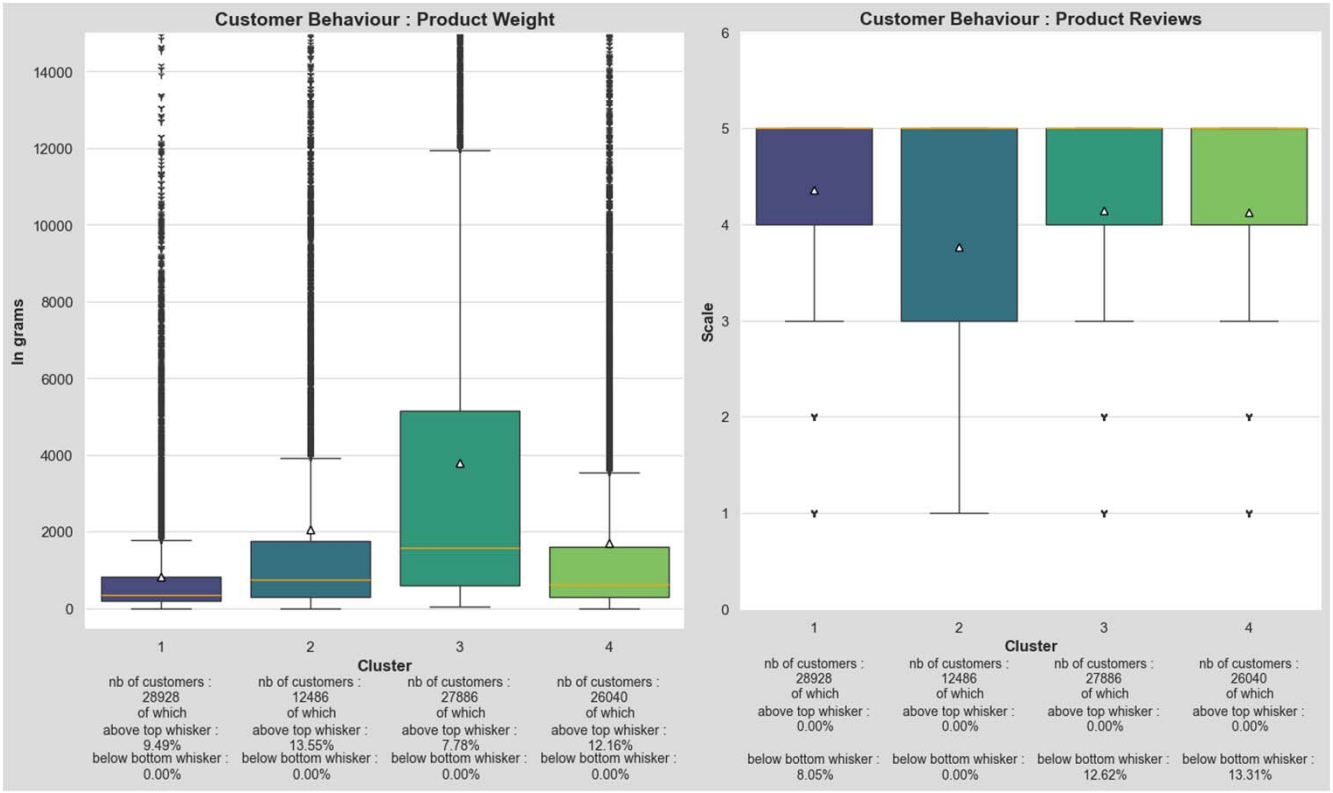
Les 5 codes postaux avec plus de 30 reviews enregistrant le

pire score moyen

sur les 12 derniers mois sont tous localisés à

Rio de Janeiro

Annexe 4a – Autres Indicateurs – Comportement Clients

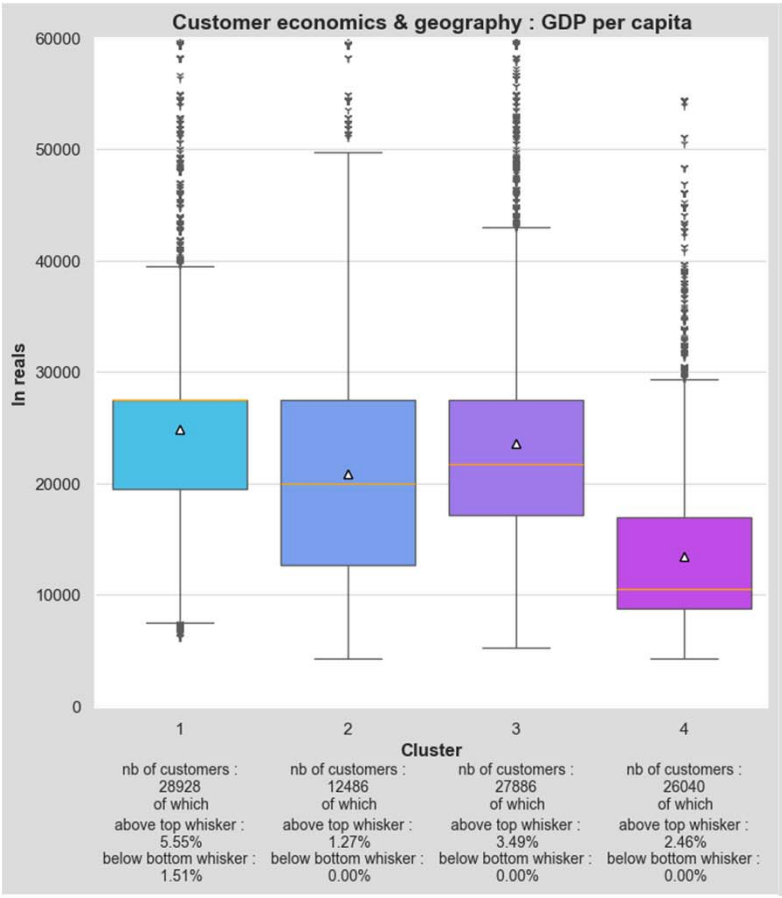


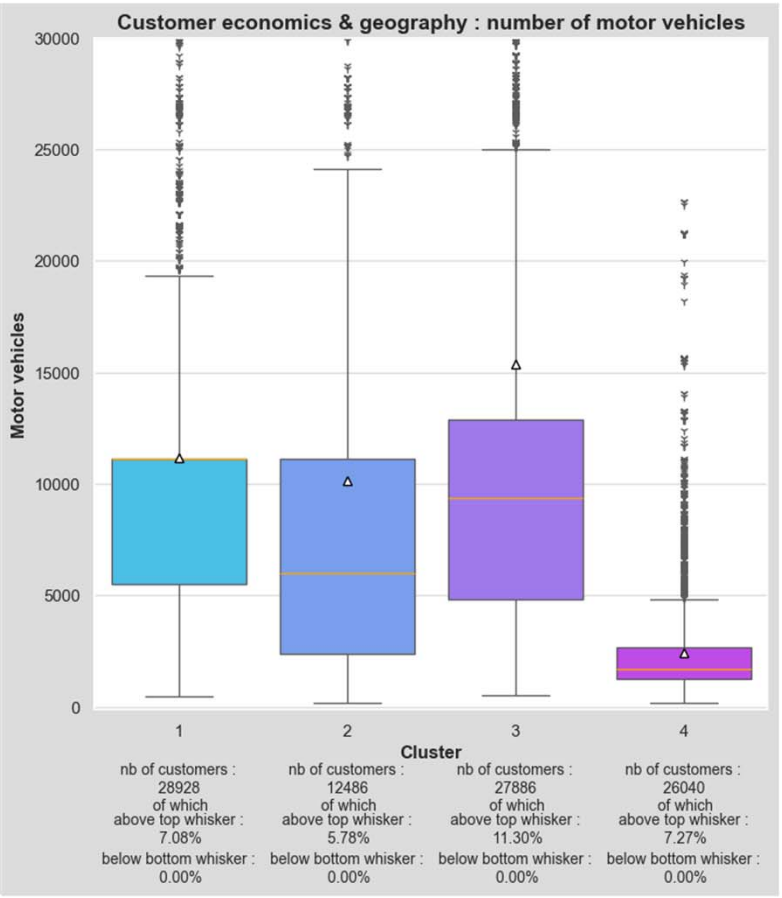
Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives entre tous les clusters sauf:

- <u>Product reviews:</u>

Clusters 3 et 4

ANNEXE 4B – AUTRES INDICATEURS – ÉCONOMIE & GÉOGRAPHIE*



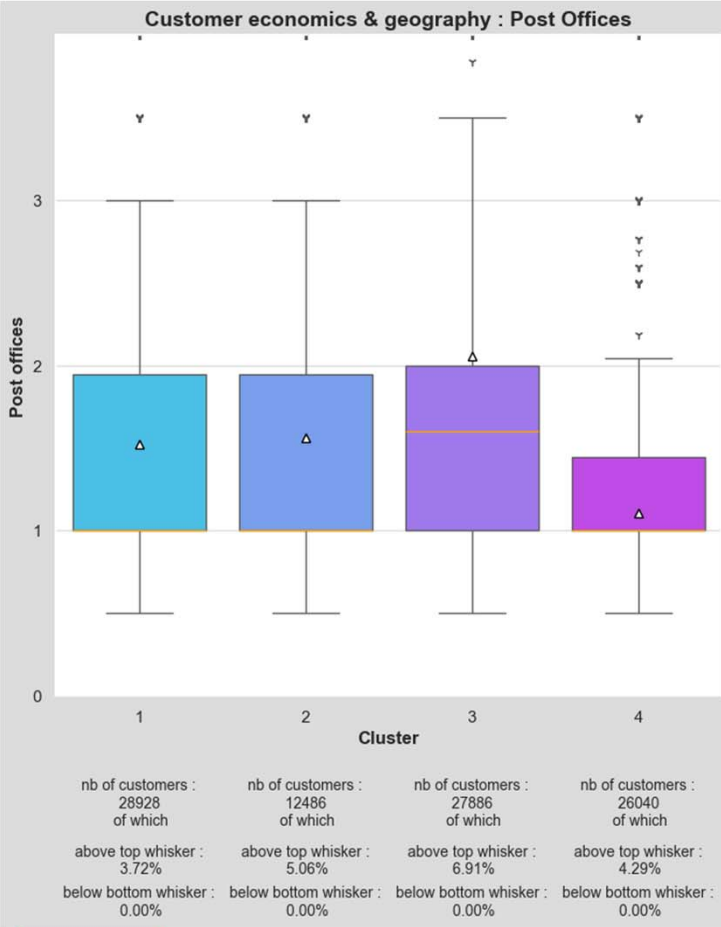


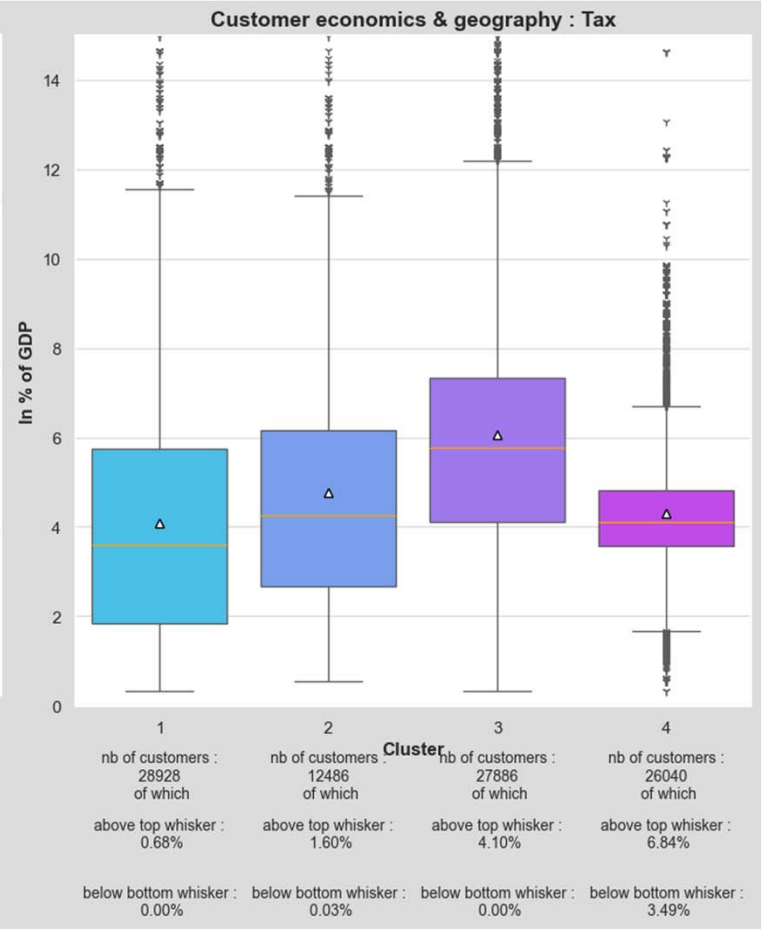
Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives entre tous les clusters sauf:

- <u>Motor vehicles :</u> Clusters 1 et 3

*Source : données des recensements de la population Brésilienne 2010 & 2022 disponibles ici https://www.kaggle.com/datasets/crisparada/brazilian-cities

ANNEXE 4B – AUTRES INDICATEURS – ÉCONOMIE & GÉOGRAPHIE*

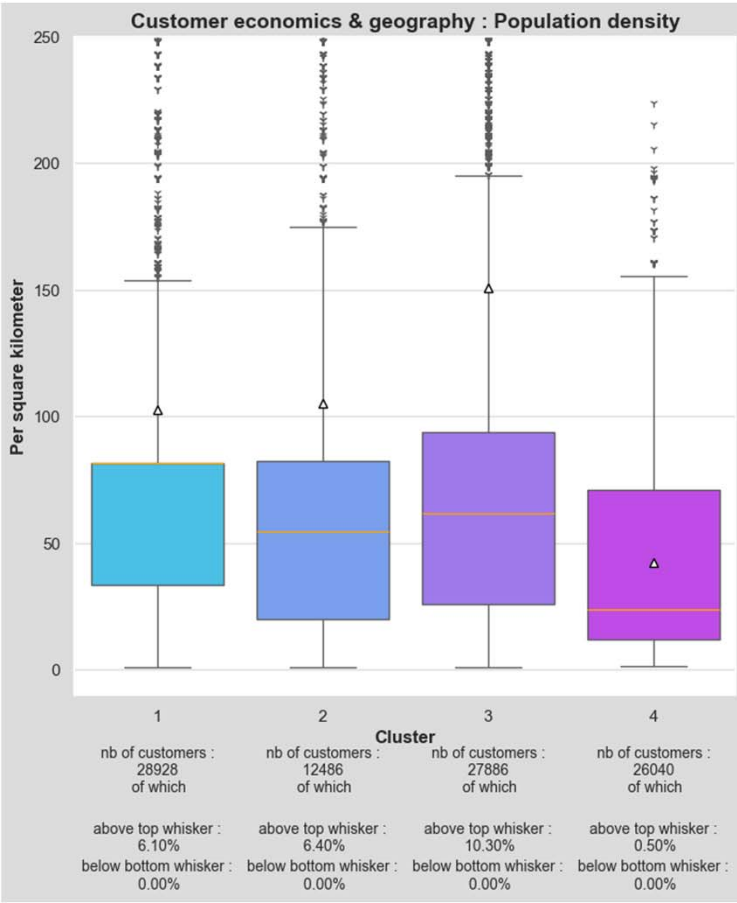


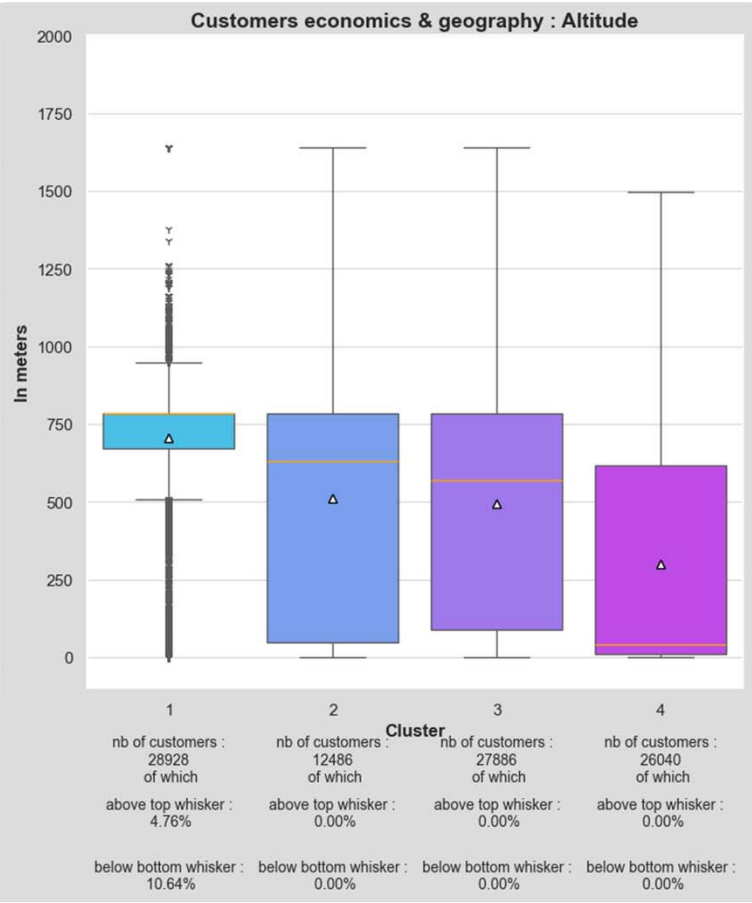


Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives entre tous les clusters.

*Source : données des recensements de la population Brésilienne 2010 & 2022 disponibles ici https://www.kaggle.com/datasets/crisparada/brazilian-cities

ANNEXE 4B – AUTRES INDICATEURS – ÉCONOMIE & GÉOGRAPHIE*



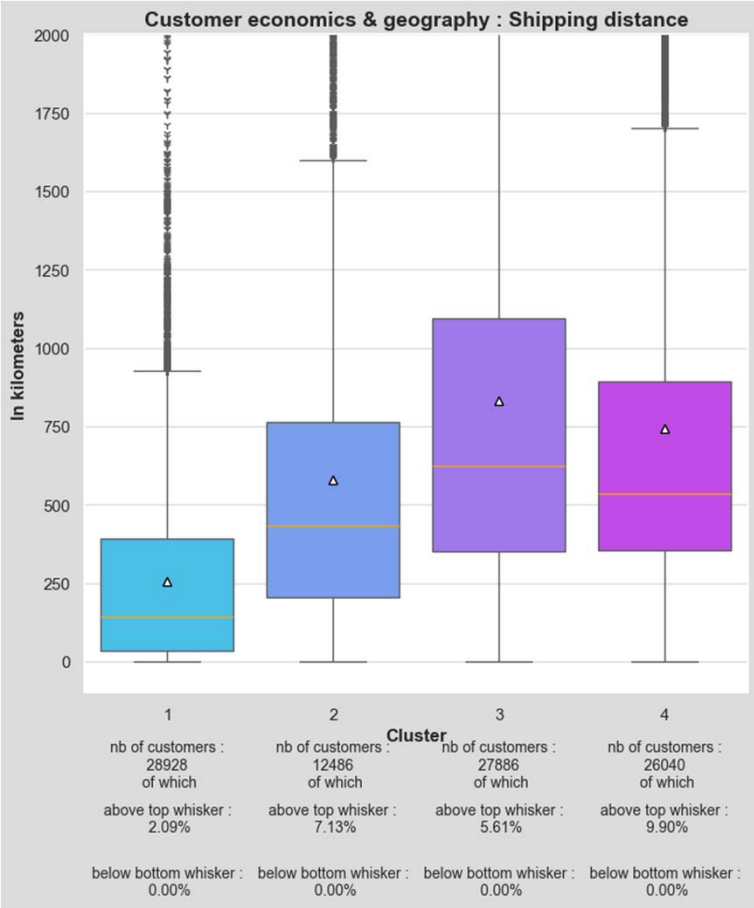


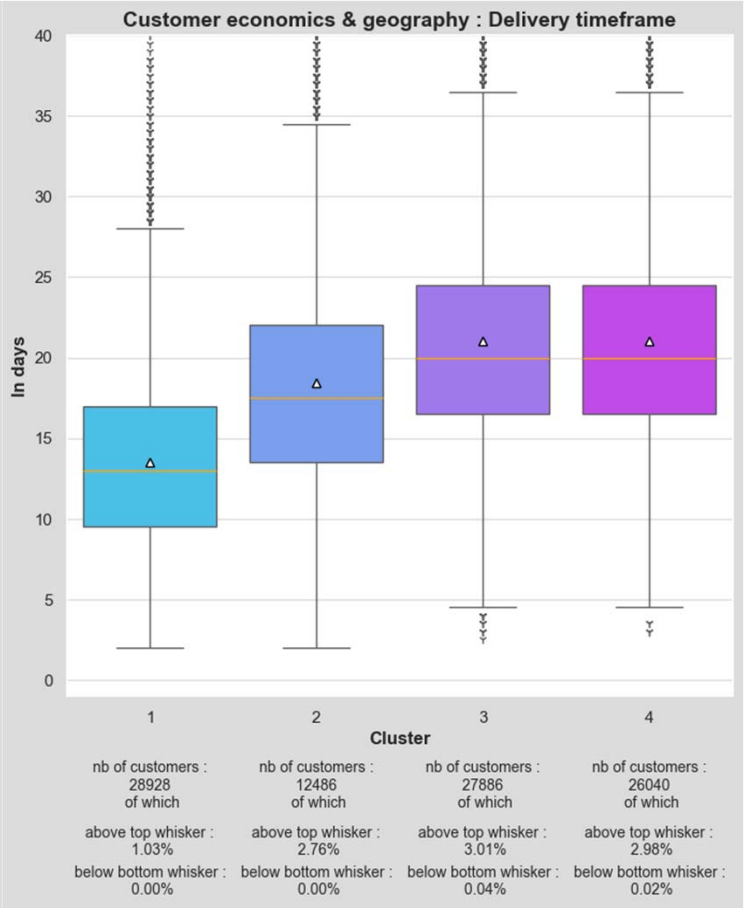
Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives entre tous les clusters sauf:

- <u>Product reviews:</u> Clusters 3 et 4

*Source : données des recensements de la population Brésilienne 2010 & 2022 disponibles ici https://www.kaggle.com/datasets/crisparada/brazilian-cities

ANNEXE 4B – AUTRES INDICATEURS – ÉCONOMIE & GÉOGRAPHIE*





Tests de Kruskal-Wallis & Dunn => différences statistiquement significatives entre tous les clusters sauf:

- <u>Delivery timeframe:</u>

Clusters 3 et 4

*Source : données des recensements de la population Brésilienne 2010 & 2022 disponibles ici https://www.kaggle.com/datasets/crisparada/brazilian-cities

Annexe 4c – Segmentation – Critères Additionnels - Synthèse

Cluster 1 – en villégiature

CSP++, zones à forte densité de population en altitude, délais de livraison plus longs malgré des distances plus courtes, accès moindre aux infrastructures mais forte densité de véhicules à moteur

Cluster 2 - citadins

CSP+, davantage de véhicules et meilleurs accès aux infrastructures que les banlieusards

Cluster 3 – banlieusards

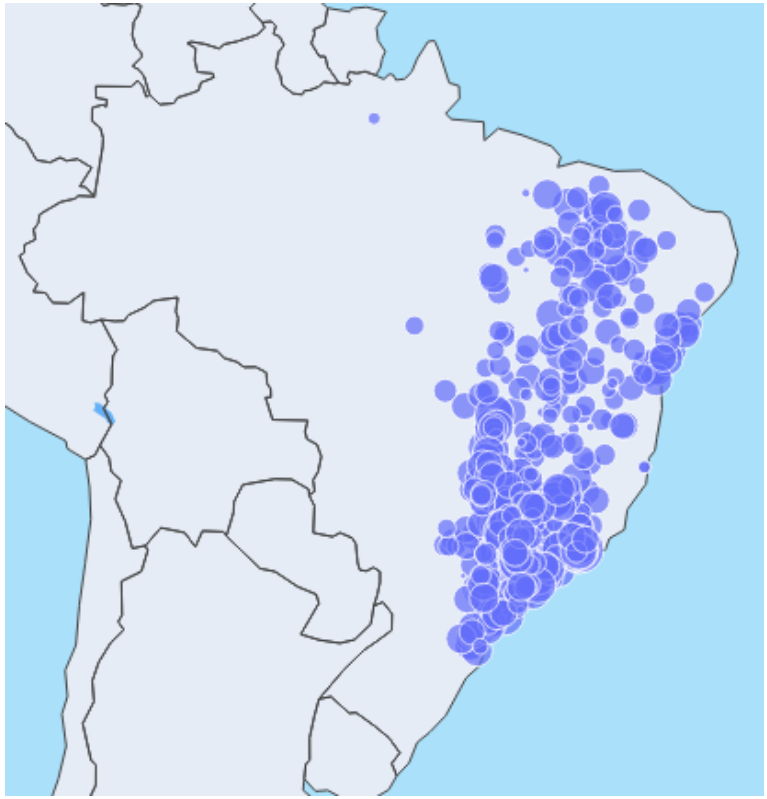
CSP+, bon accès aux infrastructures, gros achats lourds / volumineux, délais de livraison longs, review moyenne la plus basse

Cluster 4 – péri-urbains & ruraux

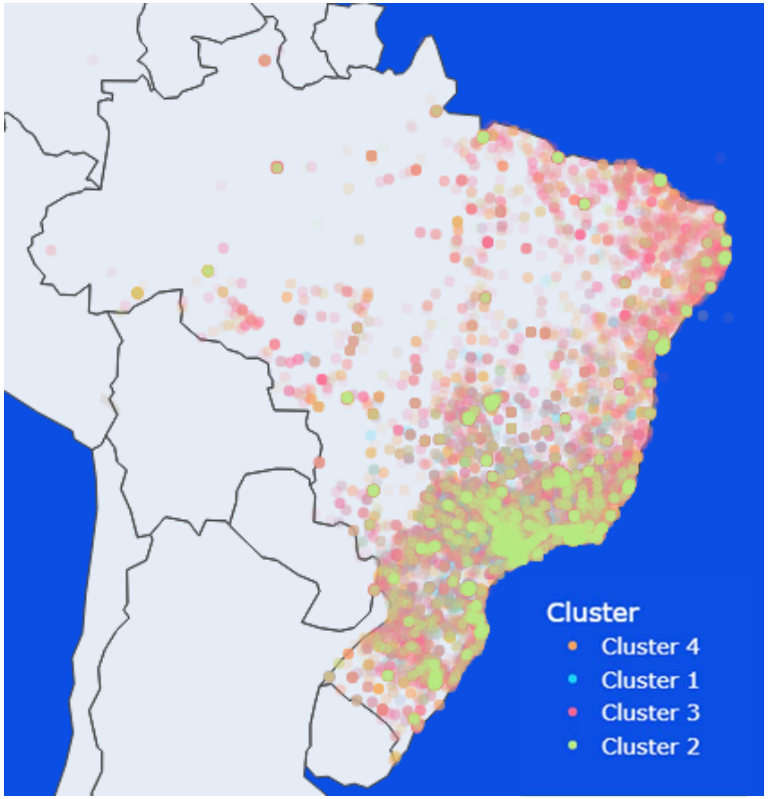
Zones à faible densité de population et basse altitude, peu d'accès aux infrastructures, CSP--, délais d'acheminement longs, grandes distances de livraison

Annexe 5 – Customers Location

MiniBatch K-Means K=1,000:



K-Means K=4 :



ANNEXE 6 – ANALYSE EN COMPOSANTES PRINCIPALES